Run II Computing

Director's Review of Computing

June 4-6, 2002

Wyatt Merritt

- **Scale of Run II Computing**
- **Status of Run II Computing**
  - **Computing systems**
  - **Software infrastructure**
  - **Reconstruction and simulation**
  - **Data handling**
  - **Offsite computing**
- **How we got here**
- **The next steps**

Scale of Run II Computing

- ***Scope of current computing:*** **Approximate!**

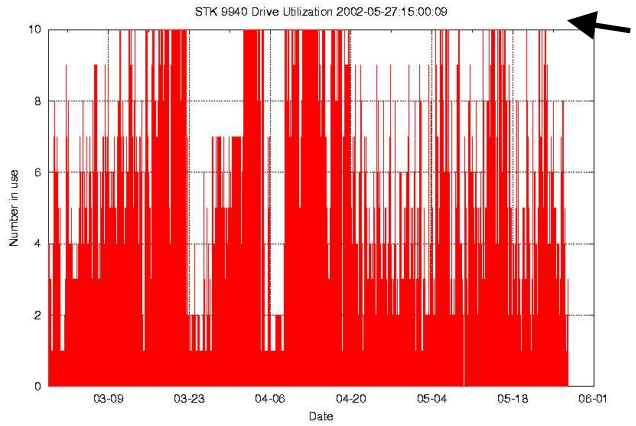
  2 experiments @ ~ 12 MB/sec each, data rate from online system
  ~ 12 MB/sec into reconstruction farms
  ~ 4 - 16 MB/sec out of reconstruction farms
  ~150MB/sec each total offline capacity for data movement

  Raw data ~ 150 TB /yr /expt
  Total datasets ~ 500 TB /yr /expt (including raw, reconstructed, derived and simulated data)
  Central disk storage ~ 30 TB /expt (growing!)
  Offline production CPU ~ 40000 SpecInt2000 (growing!)
  Offline analysis CPU ~ 1500 SpecInt2000 (increasing rapidly!)

- **Both experiments logging data reliably and moving data in and out of mass storage on a scale well beyond Run I capability (several TB's / day)**
  Both experiments reconstructing data approximately in real time with reasonable output for start-up analyses
  Both experiments providing analysis CPU to 150-300 users/day

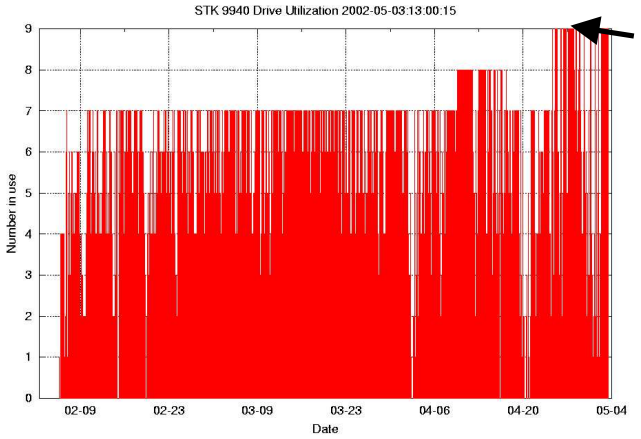
Scale of Run II Computing

CDF STK drive utilization



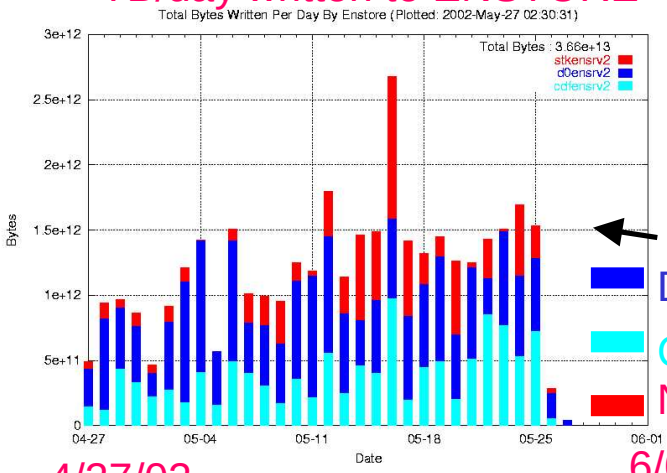
10 drives

DØ STK drive utilization



9 drives

TB/day written to ENSTORE



1.5 TB

DØ

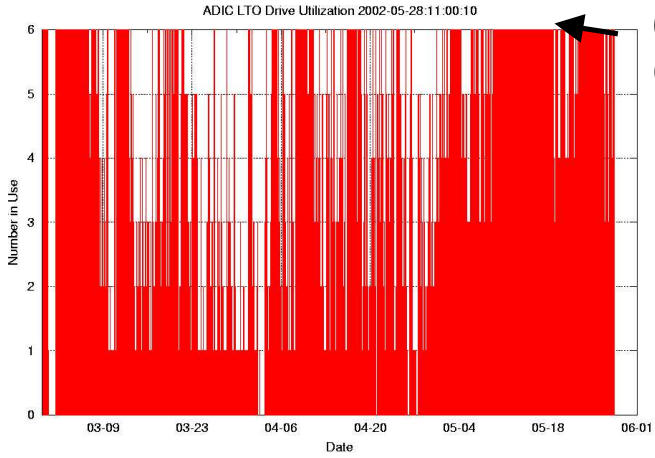
CDF

Non-RunII

4/27/02

6/01/02

DØ LTO drive utilization



6 drives



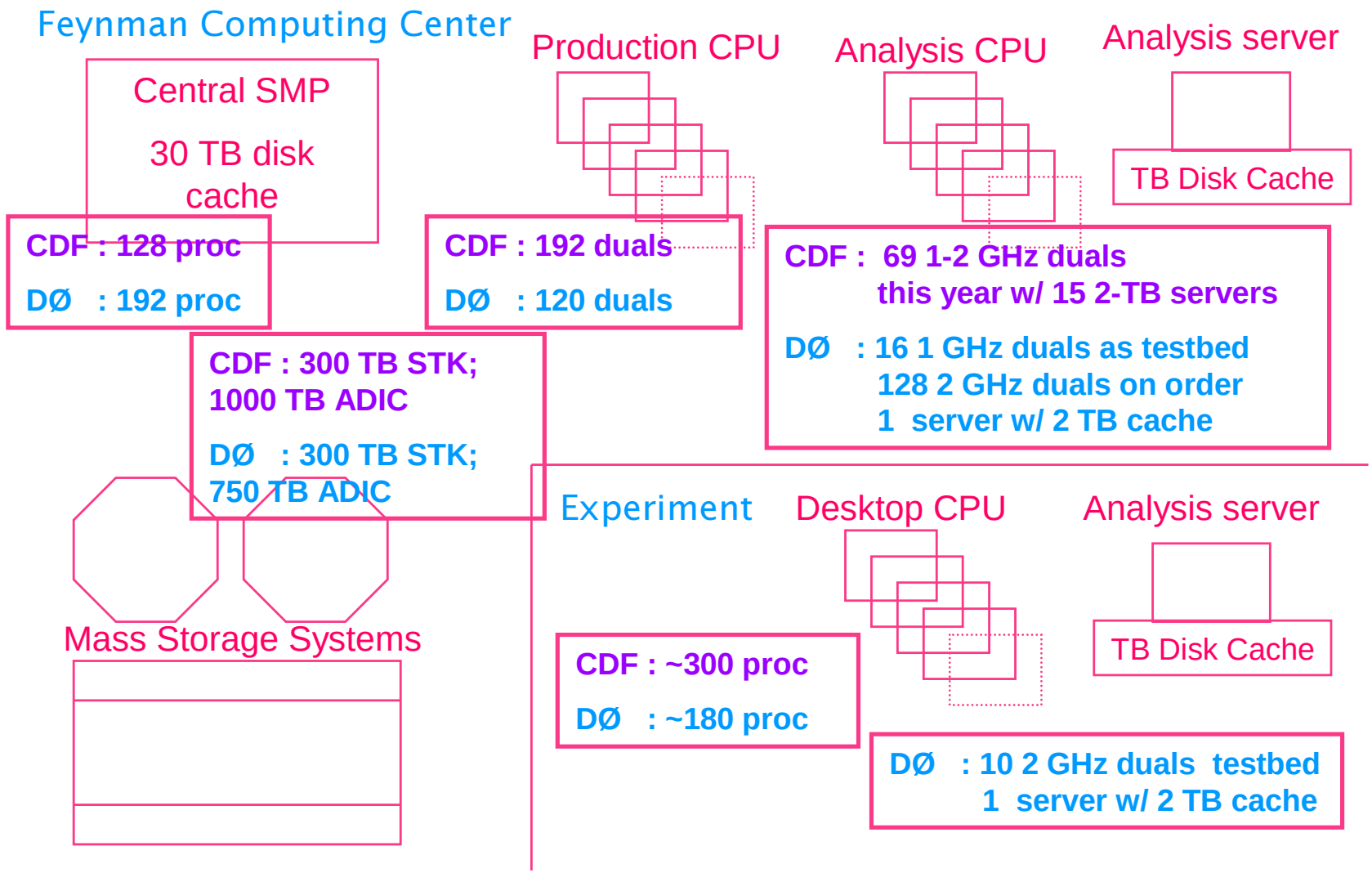





Computing Systems for Run II

- Both experiments can currently be described by the same basic model
  - Robotic mass storage system as primary repository
  - Large SMP with high I/O capacity and disk throughput
  - Linux farms for production activities with high CPU capacity
  - Linux analysis CPU - with several directions available based on working model and support model
- Differences in detail
  - DØ deploys two types of tape drive in production
  - CDF is in the process of deploying TB Linux disk servers on a large scale
- Future plans may exercise different optimizations, at least in the short term
  - CDF plans to concentrate on a central analysis farm implementation
  - DØ is exploring SAM servers located at the experiment and the use of GRID tools to leverage offsite resources -- a more distributed system
- **We can use each others' experience as backup and augmentation !**
  - cf. CDF's adoption of ENSTORE & RCP & possible adoption of SAM, and DØ's watching with interest the deployment of Linux TB file servers on a large scale

Computing Systems for Run II



Software Infrastructure

- **Databases** - based on ORACLE
  - Both experiments pleased with excellent DB administration from the CD: the ORACLE DBs and DB server machines have maintained high availability. Event / File catalogs, Run & Luminosity DBs, Calibration DBs, Trigger DBs operational for each experiment. Plenty of work remains in writing, debugging, and tuning DB applications for the experiments!
- **Common code** – the Fermilab C++ class library (ZOOM) & CLHEP
  - Includes histogramming interface, physics vectors, linear algebra, a parameter input package called RCP, etc. Also used by other expt's!
- **Code Management** - based on CVS (freeware) and SoftRelTools (Fermilab)
  - Used by BTeV, CDF, DØ
  - Each Run II experiment makes tiered releases (development and production) according to a regular schedule
- **Compilers and Debuggers** - common choice of CD-supported products
  - KAI C++ compiler currently in use, but must be phased out by late '03 (no vendor support available). CD & both experiments testing gcc (freeware).
  - Totalview & gdb the debuggers of choice; insure and purify leak-checkers also available to the experiments

Infrastructure ~ continued

- **User analysis framework** - based on ROOT
  - Both experiments use ROOT (freeware with a support arm in the Fermilab CD) as their tool for end-user analysis (making ntuples and histograms) [CDF also uses ROOT I / O as its persistent data format.]
- **Simulation code** - based on common set of physics generators and on the GEANT3 detector simulation (although each experiment has its own fast parametrized simulation programs)
- **Security** - both experiments' systems are fully Kerberized in accordance with the Laboratory computing security plan.
  - CDF and DØ were the first experiments to implement closed Kerberos systems at the lab.

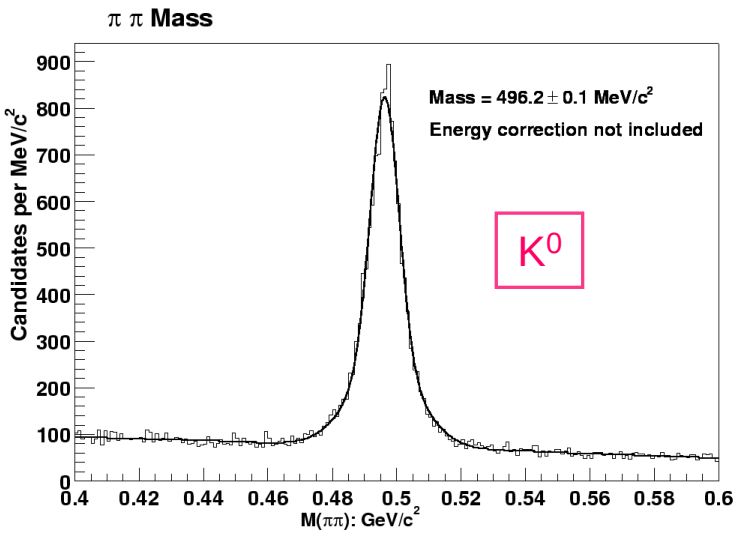
Almost all the infrastructure choices are common to the two experiments!

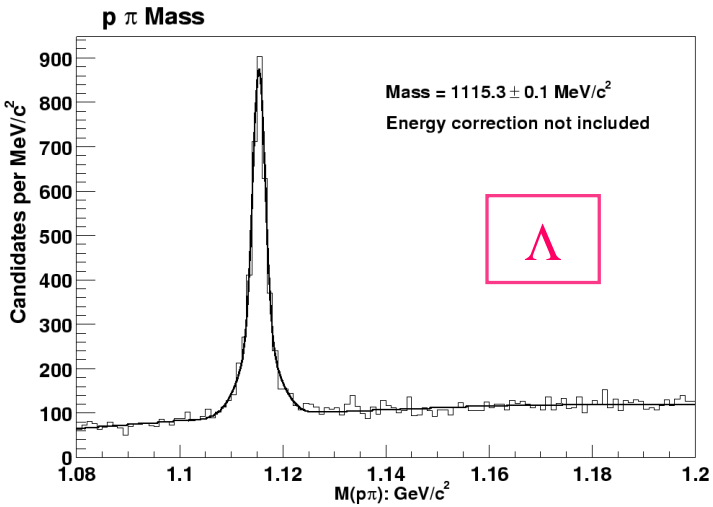
This has been a successful effort to maximize the support benefits from the fixed CD resources available.

Reconstruction & Simulation - CDF

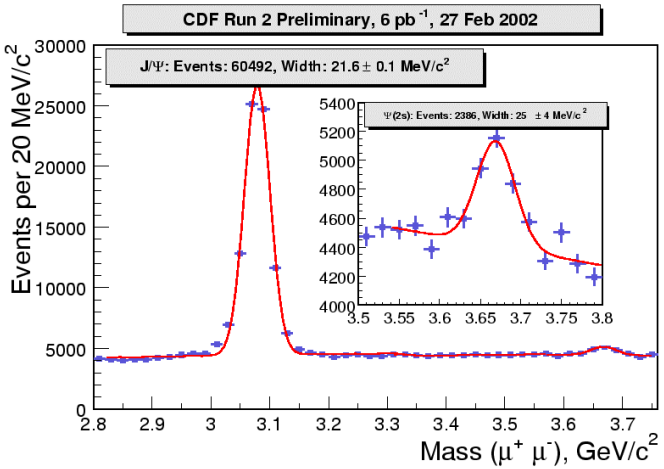
- Reco executables routinely constructed and tested
- ~ 5 sec/evt on PIII 500 MHz machine
- Stable software for jets, electrons, photons, muons, taus, and tracks.
  - Jets, electrons, photons advanced: studying calorimeter energy scales
  - Muons: quite adequate for reconstruction of J/ψ , W, Z
  - Taus: clean 1- and 3-prong signals from W $\rightarrow \tau$ decays
  - Tracks: working in COT drift chamber and extrapolating from COT into silicon detector
  - Tracks in silicon detector: still developing rapidly
    Works well in simulation.
    Improvements for running on real data in progress.
    Missing silicon layers, misalignment, noise, & calibrations complicate operations.

Reconstruction & Simulation – CDF

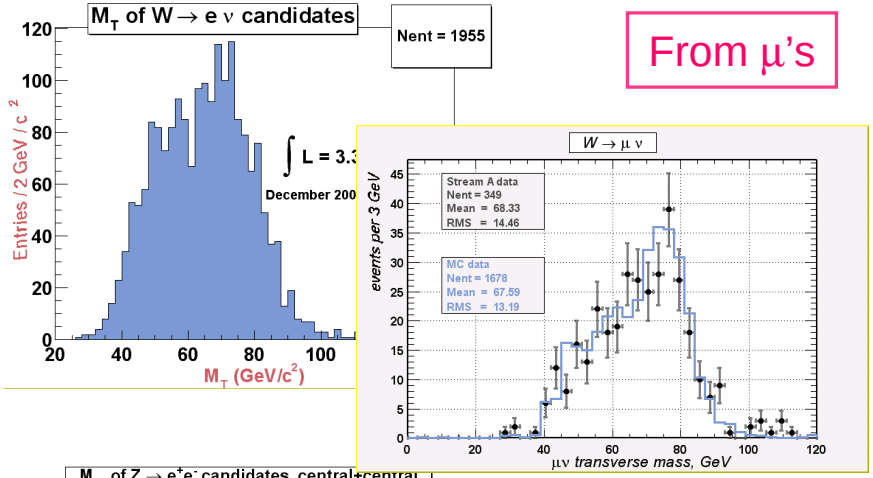




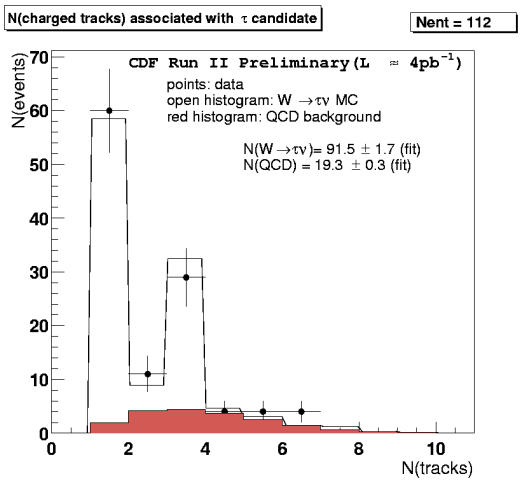
J/ ψ &
ψ(2S)

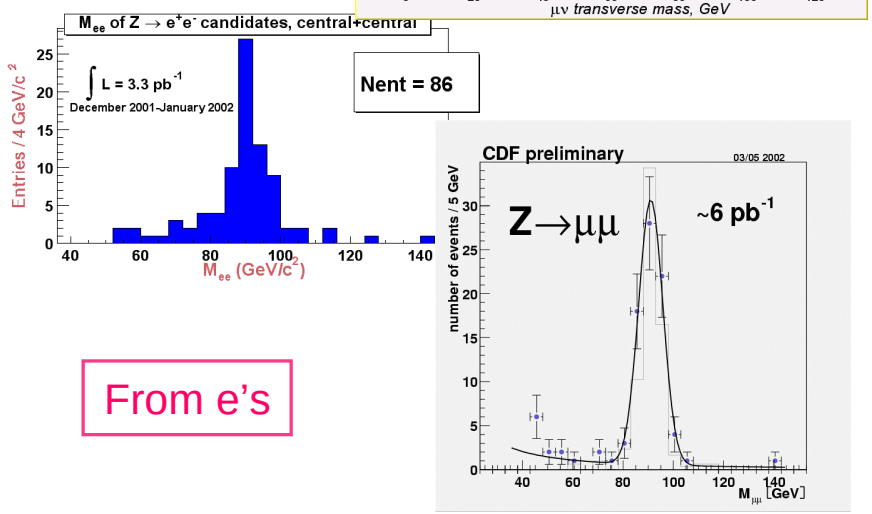


Reconstruction & Simulation – CDF
W's and Z's

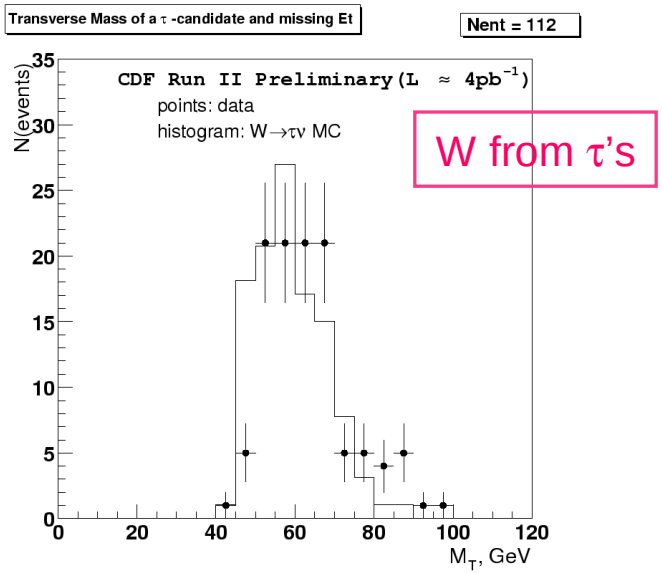


From μ's



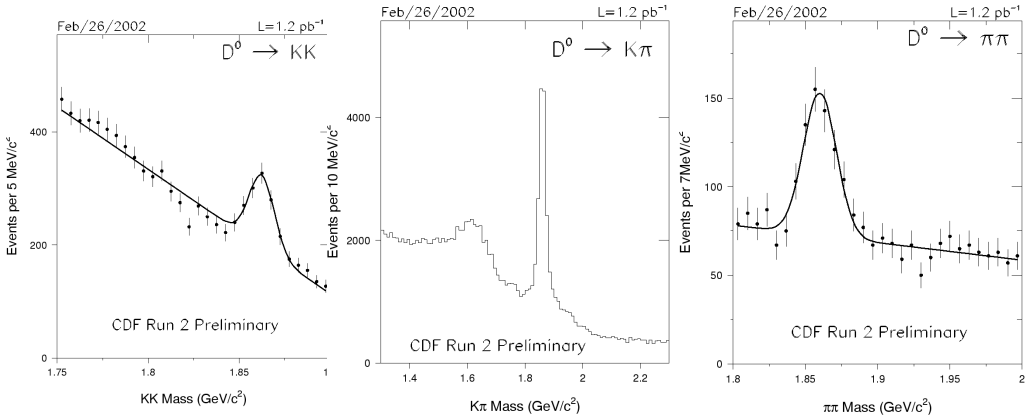


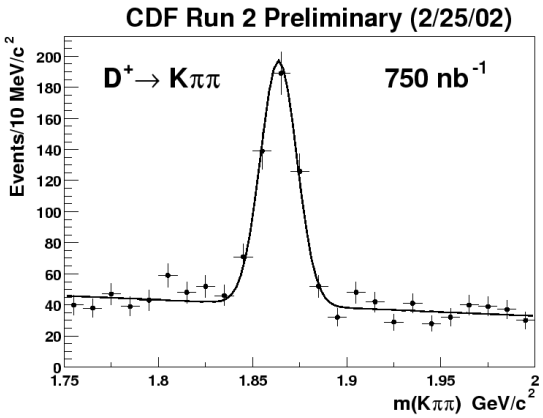
From e's

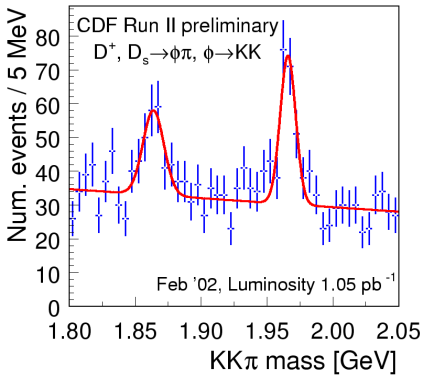


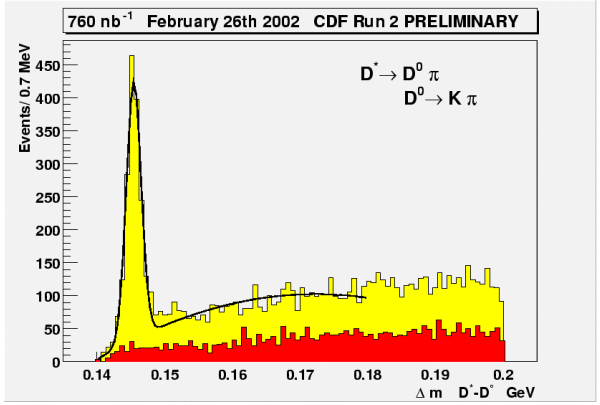
W from τ's

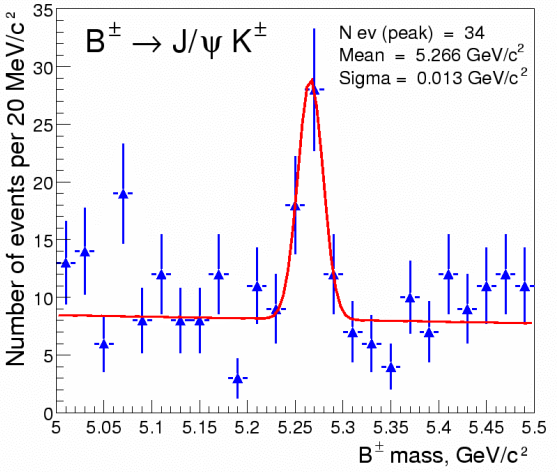
Reconstruction & Simulation – CDF
D and B mesons



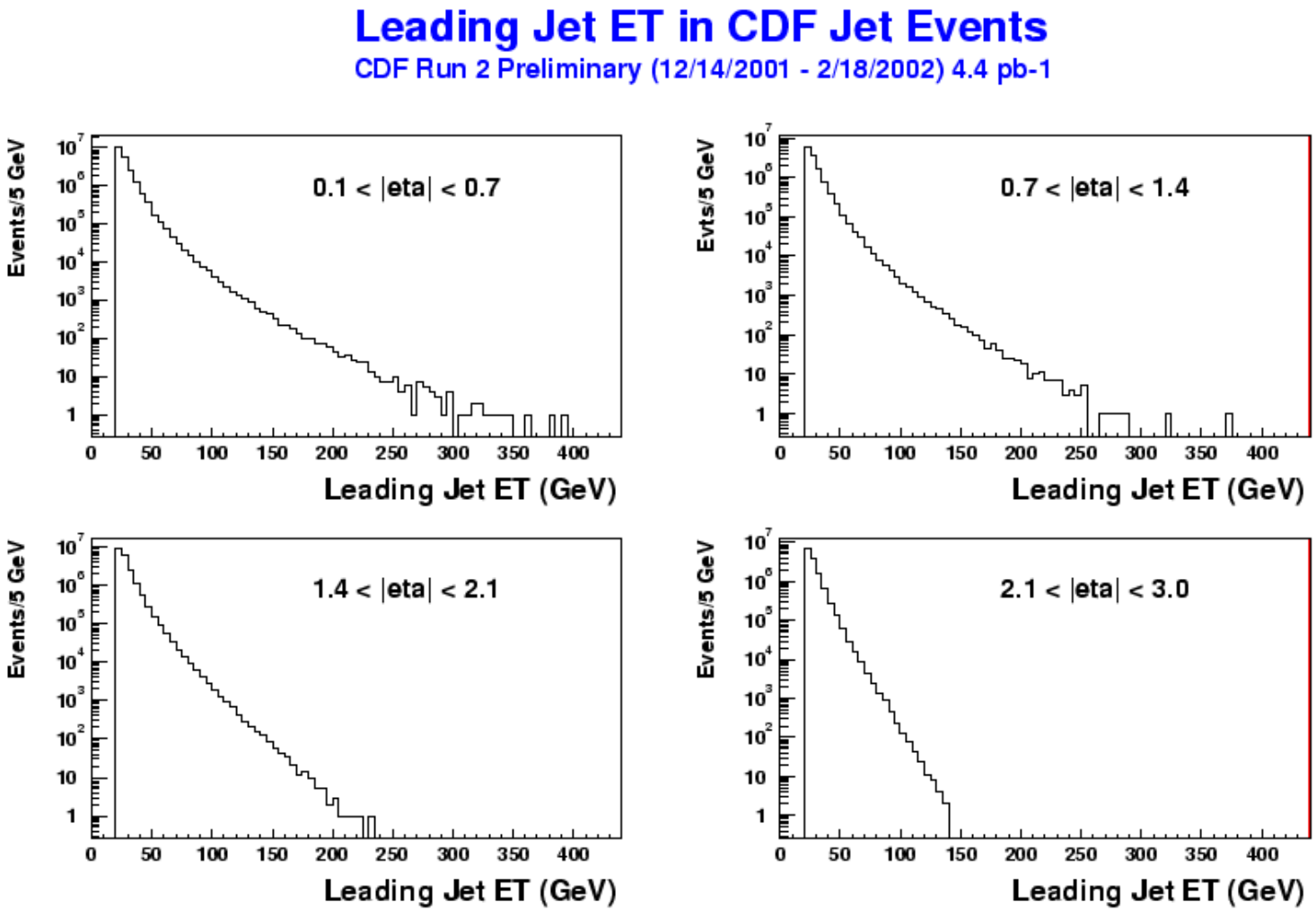






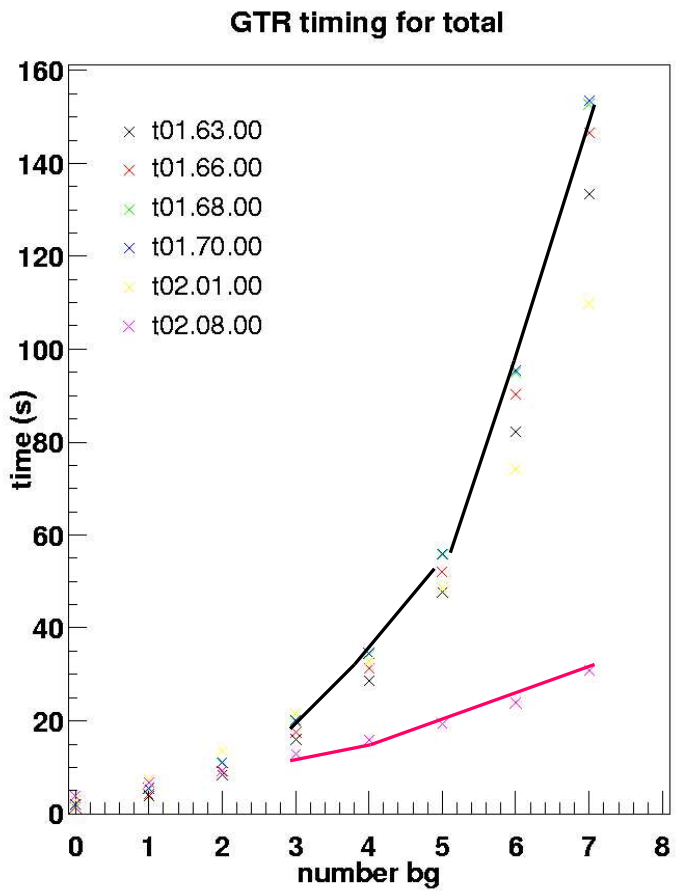


Reconstruction & Simulation – CDF



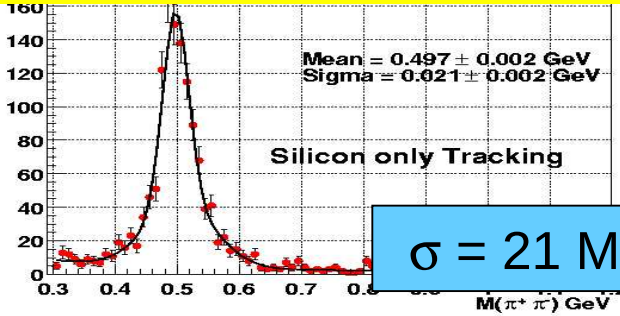
Reconstruction & Simulation - DØ

- Production pass releases are keeping up with the changing demands from the detector commissioning
- ~ 12 sec/evt in current production release (rising in next one!)
- Particle ID in place for jets, electrons, photons, muons, taus
- Tracking implemented for silicon stand-alone, fiber tracker stand-alone, and global tracks (using both tracking detectors).
- Focus on alignment, calibration, and tracking code performance
- Simulation chain operating well :
  Plate-level GEANT sim → digitization
  → bkg & noise overlay



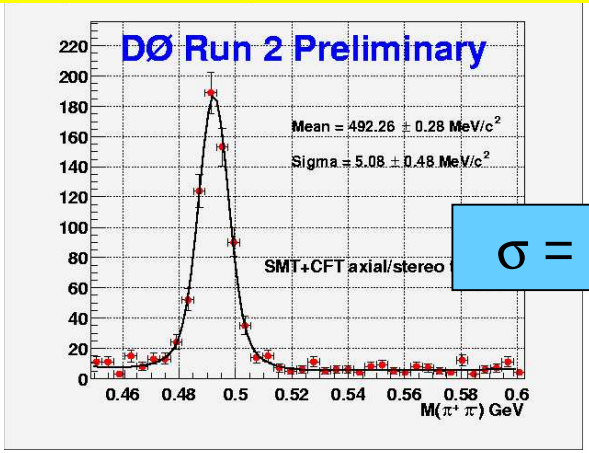
Reconstruction & Simulation - DØ Tracking

K^0 signal, silicon standalone tracking

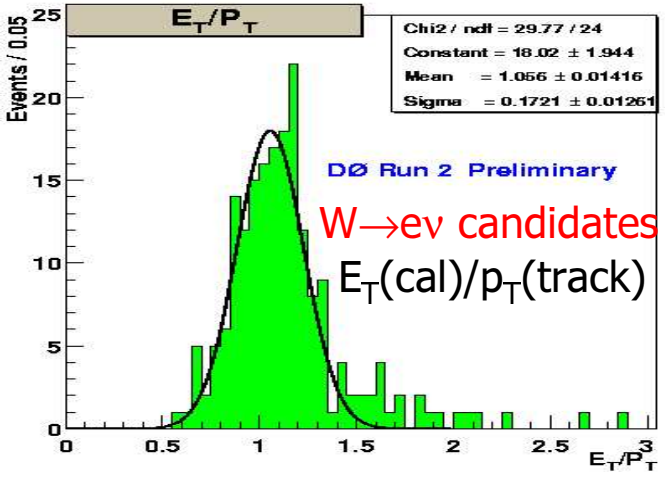


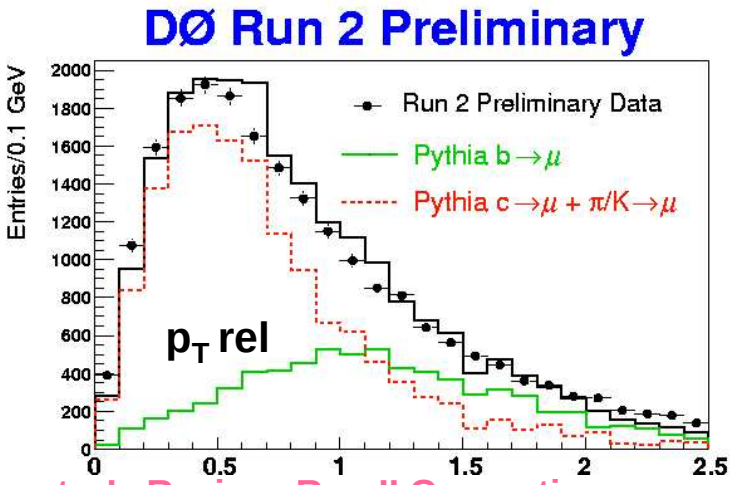
σ = 21 MeV

K^0 signal, silicon + CFT axial/stereo tracking

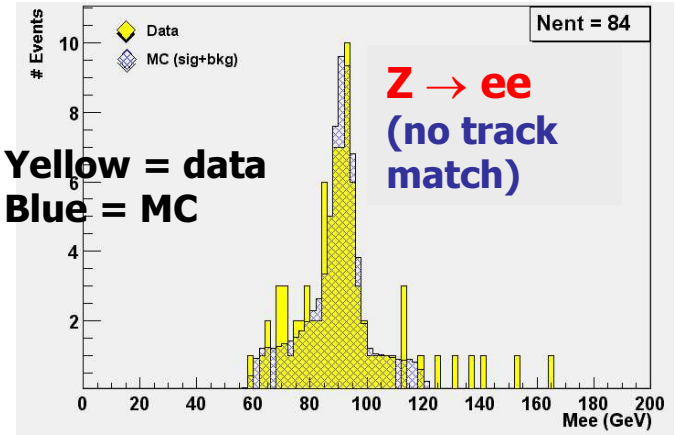


σ = 5 MeV

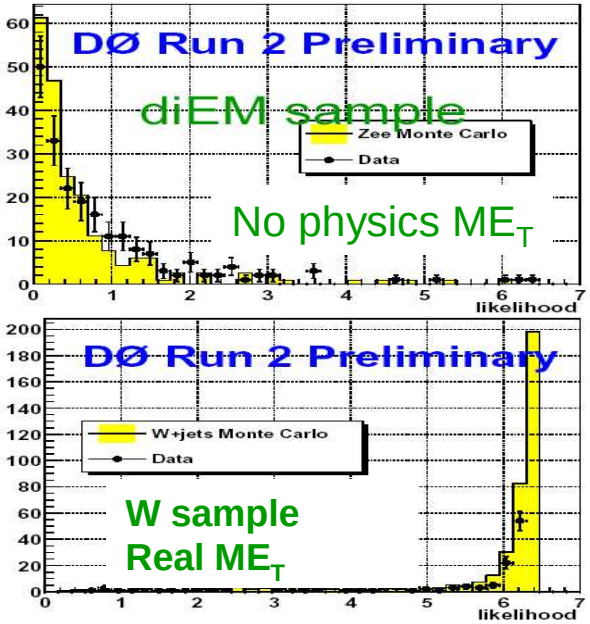


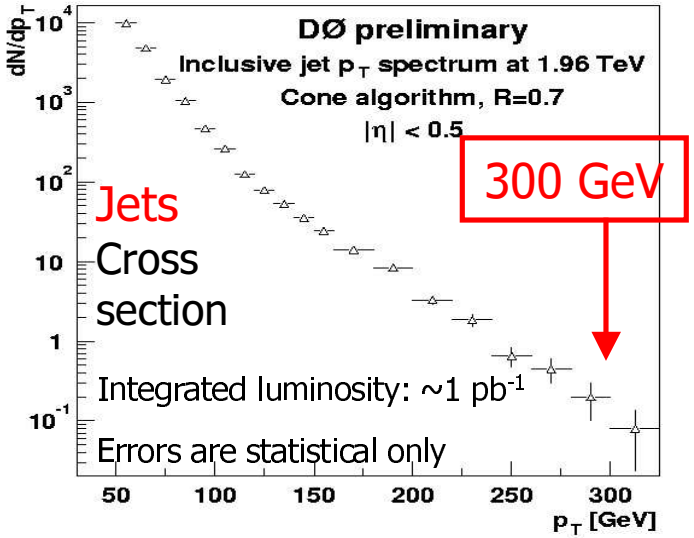


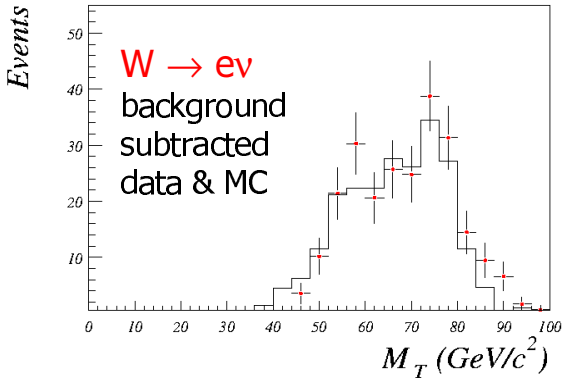
Reconstruction & Simulation - DØ Calorimetry



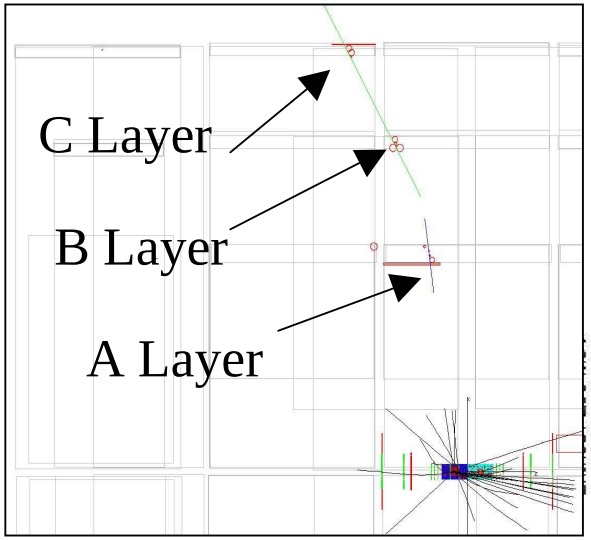
ME_T Significance is well described by Monte Carlo
→ *we understand the resolutions*



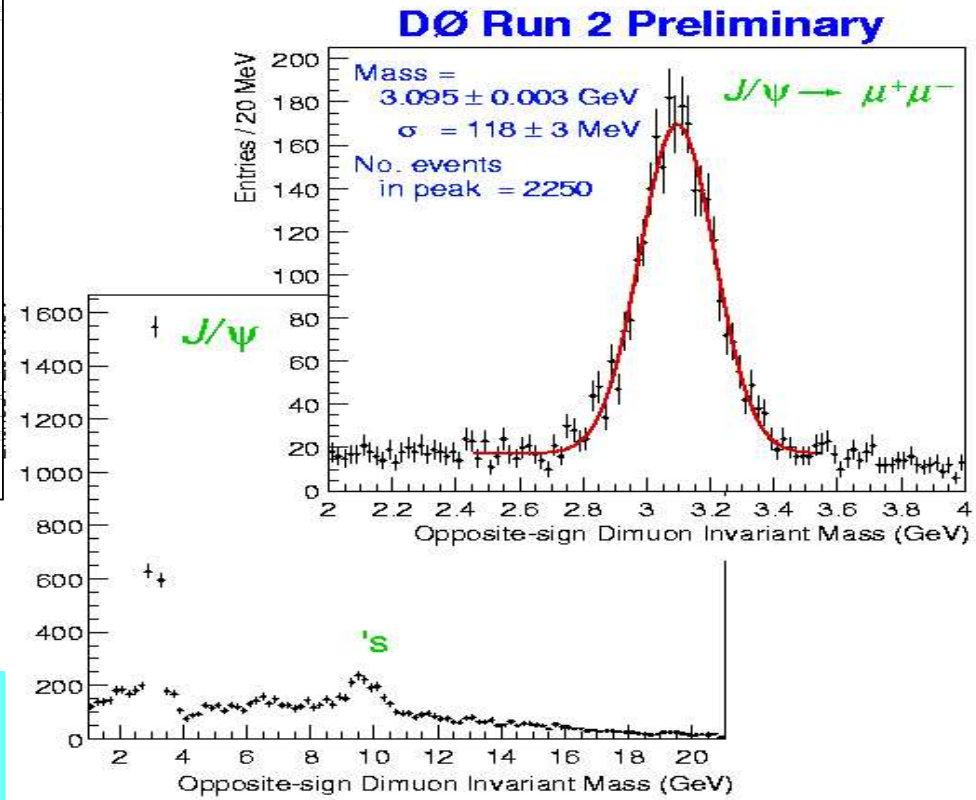




Reconstruction & Simulation - DØ
Muons



CFT Tracks matched to Muon System Tracks



- For J/ψ , if only take tracks with both silicon and CFT hits, mass resolution ~70 μm, c.f. 50–60 μm expected from Monte Carlo

Data Handling

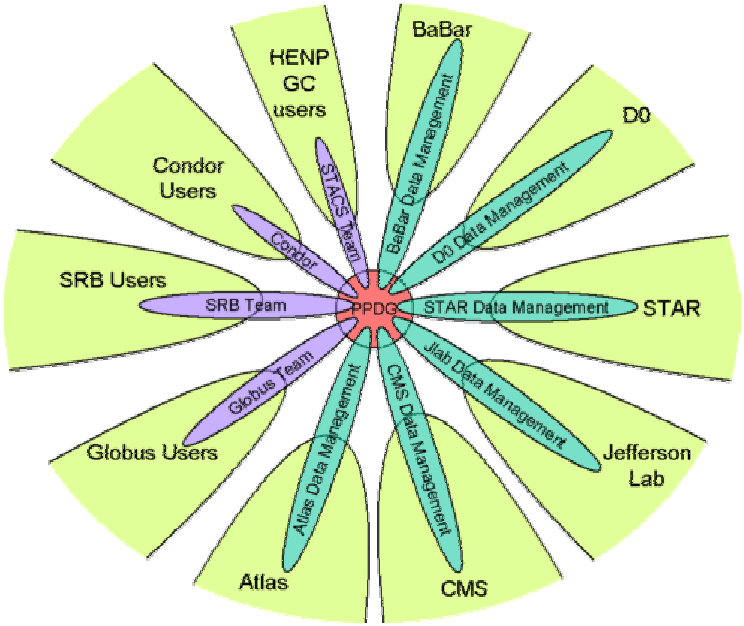
- DØ is using the SAM data handling system with the ENSTORE mass storage system, both systems developed within the Fermilab CD.
- CDF has recently adopted ENSTORE as its mass storage system, and is in the process of prototyping SAM as a replacement for their current custom data handling system
- Over the previous 6 months, both experiments switched to using STK robots and STK 9940 drives, with very significant improvement in operational reliability.

  DØ : raw datasets are written to STK robots. Using the ADIC robot with IBM LTO drives, to store its reconstructed, derived, and simulation data sets (and to test LTO drives as possible money-savers for the remainder of the Run II storage deployments). Data from any of the 3 robotic systems in use are transparently delivered to user applications.

  CDF: has copied older datasets to STK tapes and these are being used by physics groups. New raw and reconstructed data being written to the STK robot as well (and to the ADIC system as a failsafe). Also delivers data transparently to users from different robots.

Run II & Grid Computing

- Grid computing is another name for distributed computing resources.
- It's in our future with the LHC experiments -- how does it relate to Run II?
- The CD is participating in the Particle Physics Data Grid project.
- British groups from DØ and CDF are participating in GridPP – a UK Grid project



- The SAM data handling project is testing prototype grid tools with SAM.
- This effort recently succeeded in demonstrating file transfer from SAM stations in Great Britain to the DØ central SAM station using grid security mechanisms.
- We hope to offer the grid projects realistic test setups in return for development effort on tools we need for distributed systems.

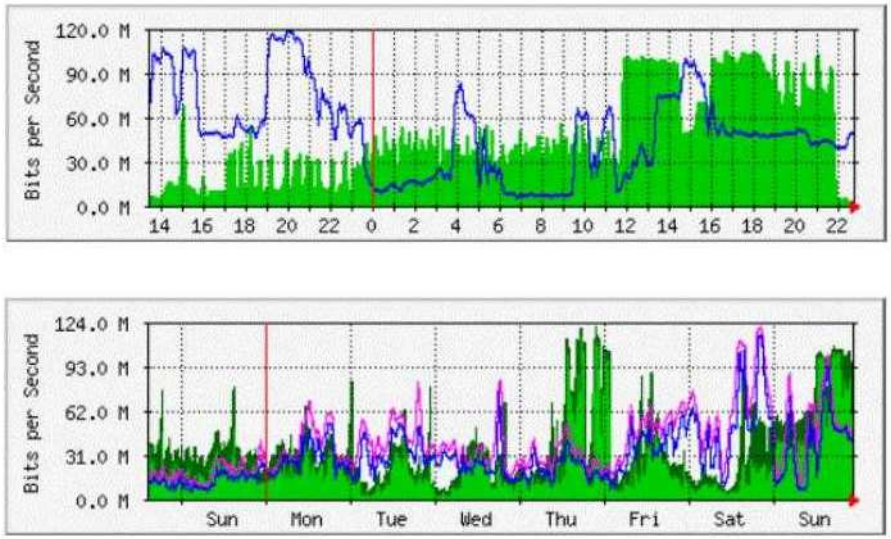
Offsite Computing

- Offsite computing 'in real time' has become an important way of working for High Energy Physics
  - **Code development**
  - **Generation of simulated data**
  - **User analysis of (subsets of) collider data**
  - **(Re-)processing of collider data**
- CDF distributes its code nightly to 30 remote institutions and makes available its express raw stream and derived datasets offsite.
- DØ has brought online 6 production farms for generating simulation data (~400 Linux & IRIX CPUs).
- DØ has also installed its data delivery system at 21 remote institutions.

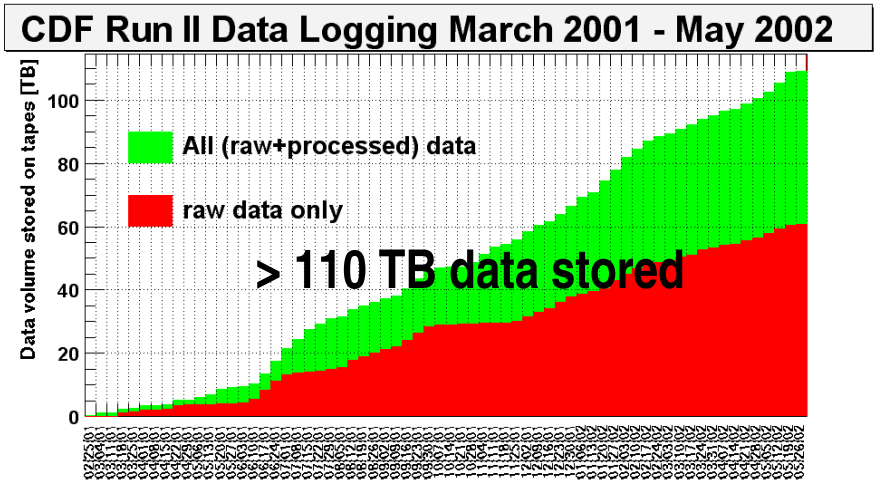
Note that heavier use of networked data delivery ~ in both directions ~ makes upgrading the Fermilab network connection to the outside world an important event in the very near future!

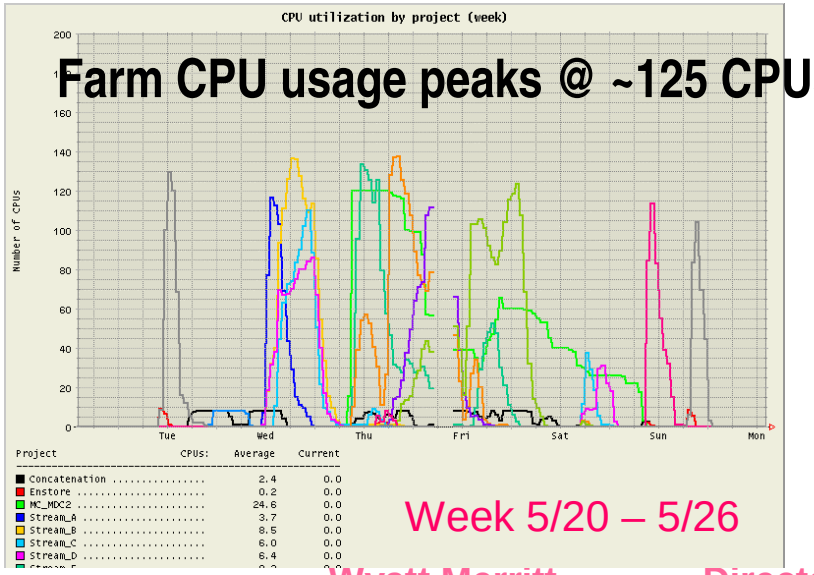
Networking Status

- Fermilab currently has 2 OC-3 connections (155 Mbits/sec each) for ESNET and MREN
- An upgrade of the ESNET connection to OC-12 (622 Mbits/sec) is planned. It requires a new fiber to the closest Qwest connection (by late summer?)
- The plots below illustrate recent tests that have saturated the site's capacity for limited periods.

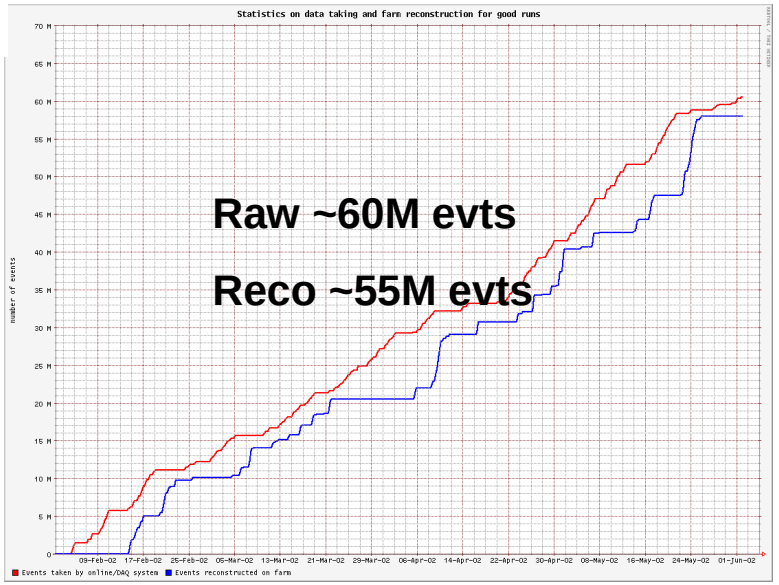


Data Handling Plots - CDF





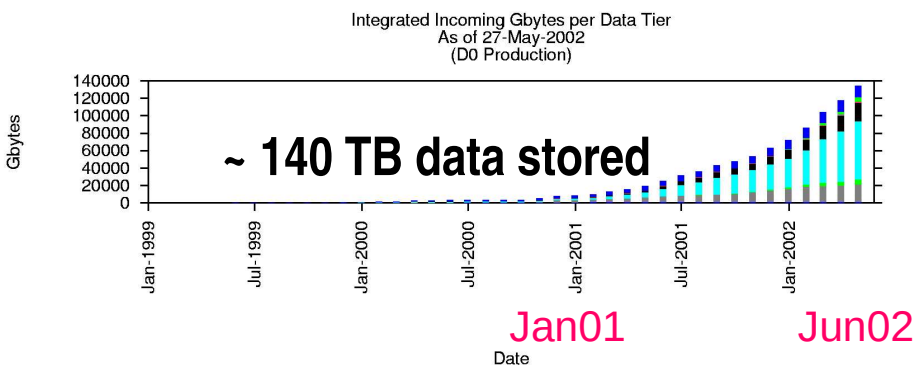
Week 5/20 – 5/26



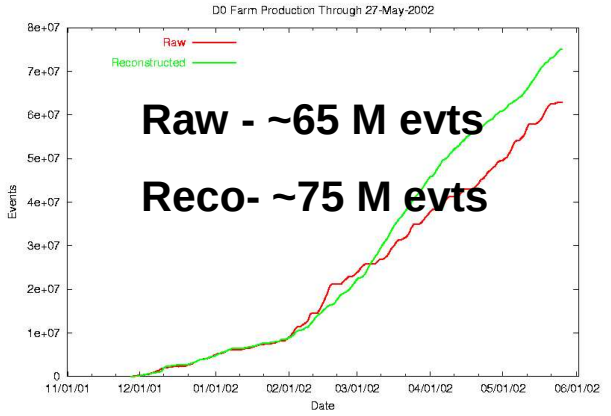
Feb 02

Jun 02

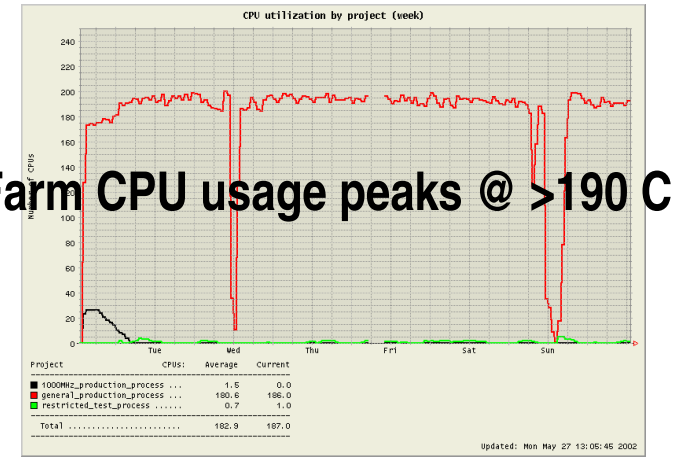
Data Handling Plots - DØ



Jan01 Jun02

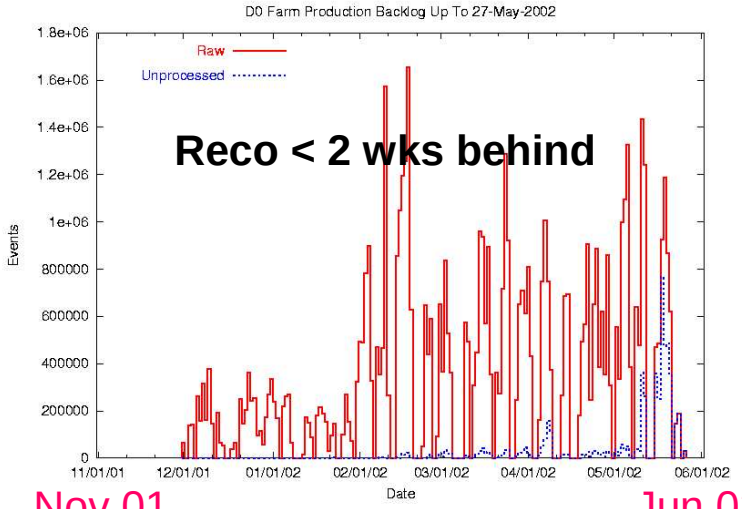


Nov 01 Jun 02



Farm CPU usage peaks @ >190 CPUs

Week 5/20 – 5/26



Nov 01 Jun 02

How We Got Here - A Bit of History

- Budget planning document June 1997
- Financial profile 1998 -- 2002 (stretched by 2 years)
- Project has coped with :
  1. One technical crisis (tape technology)
  2. Some underfunded needs (networking, server systems, infrastructure)
  3. The confrontation with real commissioning

  and will stay within the stretched profile!
- In FY02: ramping up farm and analysis systems for full load

Fiscal Year	MSS	Farms	Analysis	Disk	Other	Sum (CDF/DØ)
Spent in FY98	$1.2M	$200K	-	$200K	$400K	$2M
Spent in FY99	$2.2M	$700K	$2M	$800K	$300K	$6M
Spent in FY00	$450K	$350K	$100K	$300K	$800K	$2M
Spent in FY01	$675K	$500K	$2.1M	$600K	$300K	$4.3M
Budget FY02	*$0.8M*	*$950K*	*$1.4M*	*$500K*	*$350K*	*$4.0M*
Total	*$5.3M*	*$2.7M*	*$5.6M*	*$2.4M*	*$2.1M*	*$18.3M*
Total Needs '97 est.	*$4.8M*	*$2.8M*	*$6.4M*	*$2.6M*	*$1.6M*	*$18.2M*
Plan for FY03						*$ 4.0M*
Continuing Operations & Upgrades (FY04 and beyond)						*$4.0M*

A Bit More History

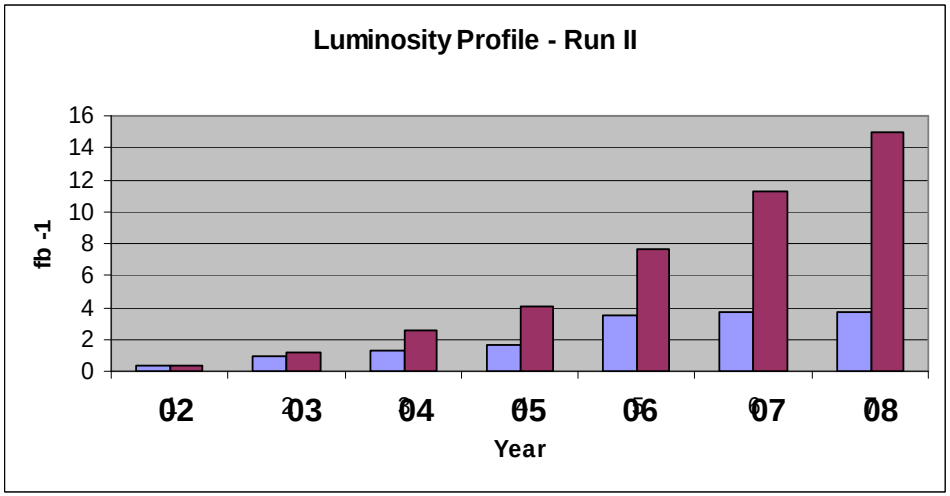
- **Joint Working Groups for Run II Computing**
  - Joint farm development & operations group – already a Fermilab tradition
  - Joint committee to choose ROOT
  - Joint procurement operations for SMP purchase
  - Joint work in the early stages of data handling
  - Framework for communication --
    - Run II Steering Committee (early days)
    - Run II Computing Operations meetings (currently)
      DØ , CDF, and CD division departments in a monthly roundtable

The Rest of Run II

- Further consolidation on data handling system is likely. What about the GRID? For distributed computing to be a serious Run II component, experiments will need to work together with CD & Directorate on making the case for expanding Fermilab's Internet access.
- C++ Working Group reconvened to look for KAI compiler replacement.
- Code sharing via the ZOOM repository continues and expands, and now extends to other experiments.
- The experiments can share information on system choices – how best to utilize large disk caches with Linux, how to get the most out of our large SMP systems, how to navigate the tape technology changes, if / when / how to move to disk storage as a primary storage medium.

The Planning Process

- Each experiment has undertaken a planning process for operations and upgrades of its Run II computing systems, involving its own collaboration and the CD departments
- Common working assumptions
  - Luminosity profile for Run II – from the Directorate
  - Cost projections – Moore's Law, but include server costs for disk projections



- Collaboration-specific assumptions
  - Data rates – from physics menus and current experience
  - Data format structures and access patterns
  - Computing system evolution – specifically SMP migration strategy (but this may become common if only one of the two paths works out, or if one becomes a clear winner !)

Conclusions

- Run II Computing had a successful planning process for the first stage of the run
- Run II Computing is functional for both experiments
- Planning for the operations and upgrade phase rests on current experience, on a number of common assumptions, and on the expectation that we can react flexibly when necessary